CLAIMS

What is claimed is:

1. A method comprising:

representing an input document image with a sequence of template identifiers to reduce storage consumed by the input document image; and

replacing the template identifiers with alphabet characters according to language statistics to generate a text string representative of text in the input document image.

2. The method of claim 1 further comprising extracting n-gram indexing terms from the text string by selecting alphabet characters in the text string that satisfy a predicate and then combining the alphabet characters to form n-grams, n being an integer.

3. The method of claim 2 wherein the n-grams are trigrams.

4. The method of claim 2 wherein the number of characters in the text string that satisfy the predicate is fewer than the number of characters in the text string.

5. The method of claim 2 wherein the predicate is a condition that substantially all selected characters follow respective spaces in the text string.

6. The method of claim 2 further comprising comparing the input document image with a plurality of symbolically compressed document images in a database based

on the n-gram indexing terms to determine whether one of the plurality of documents matches the input document image.

7. The method of claim 6 further comprising determining whether the one of the plurality of document images is a confidential document if the one of the plurality of document images matches the input document image.

8. The method of claim 7 further comprising prompting a user for authorization before operating on the input document image if the one of the plurality of document images matches the input document image and is a confidential document.

9. The method of claim 6 further comprising determining whether the one of the plurality of document images forms a sub-portion of an encompassing document image if the one of the plurality of document images matches the input document image.

10. The method of claim 6 further comprising prompting the user to select between hardcopy output of the one of the plurality of document images or the encompassing document image if the one of the plurality of document images matches the input document image and forms a sub-portion of a larger document image.

11. The method of claim 6 further comprising determining whether the one of the plurality of document images is copyrighted if the one of the plurality of document images matches the input document image.

12. The method of claim 1 wherein replacing the template identifiers with alphabet characters comprises replacing at least one of the template identifiers with an alphabet character selected according to a sequence of at least two alphabet characters selected to replace template identifiers that precede the at least one template identifier in the sequence of template identifiers.

13. The method of claim 1 wherein replacing the template identifiers with alphabet characters comprises replacing at least one of the template identifiers with an alphabet character selected according to a sequence of all alphabet characters selected to replace template identifiers that precede the at least one template identifier in the sequence of template identifiers.

14. The method of claim 1 wherein replacing the template identifiers with alphabet characters according to language statistics comprises replacing the template identifiers with alphabet characters selected according to a hidden Markov model.

15. The method of claim 1 wherein replacing template identifiers with alphabet characters according to language statistics comprises solving a substitution cipher by mapping the alphabet characters to the template identifiers based at least partly on frequency of occurrence of the template identifiers.

16. A method comprising using a hidden Markov model to solve a substitution cipher.

17. A document processing system comprising:

a deciphering module to generate a first text string based on a sequence of template identifiers in a first symbolically compressed document image and

to generate a second text string based on a sequence of template identifiers in a second symbolically compressed document image;

a conditional n-gram module coupled to receive the first and second text strings from the deciphering module, the conditional n-gram module being configured to extract n-gram indexing terms from the first and second text strings based on a predicate condition; and

a comparison module to generate a measure of similarity between the first and the second symbolically compressed document image based on the indexing terms extracted by the conditional n-gram module.

18. The document processing system of claim 17 wherein at least one of the deciphering module and the conditional n-gram module is implemented by a programmed processor.

19. The document processing system of claim 17 wherein the deciphering module generates the first text string by applying a hidden Markov model to the sequence of template identifiers in the first symbolically compressed document image.

20. The document processing system of claim 17 wherein the second symbolically compressed document image is obtained from a database of symbolically compressed document images.

21. The document processing system of claim 17 wherein the data processing system further comprises a scanning and compressing module that is configured to generate a digitized version of a source document and to perform symbolic compression of the digitized version to produce the first symbolically compressed document image.

22. The document processing system of claim 21 wherein the document processing system is a document copying system.

23. The document processing system of claim 21 wherein the document processing system is a facsimile transmission system.

24. The document processing system of claim 17 further comprising a second interface to couple the document processing system to a database of symbolically compressed document images and associated indexing terms, and wherein the conditional n-gram module is configured to store the n-gram indexing terms extracted from the second text string in the database.

25. The document processing system of claim 24 wherein the comparison module is configured to receive the n-gram indexing terms extracted from the first text string from the conditional n-gram module and to receive the n-gram indexing terms extracted from the second text string from the database via the second interface.

26. The document processing system of claim 17 wherein the measure of similarity is used to determine whether the first and second symbolically compressed document images match.

27. The document processing system of claim 26 wherein the data processing system further comprises a user interface to prompt a user to select between a decompressed version of the first symbolically compressed document image and a decompressed version of a third symbolically compressed document image that encompasses the second symbolically compressed document image if the first and

second symbolically compressed document images match, the document processing system further comprising an output module to output the decompressed version of the first symbolically compressed document image or the decompressed version of a third symbolically compressed document image according to input received via the user interface.

28. The document processing system of claim 26 wherein the data processing system further comprises a security module to determine if the second symbolically compressed document image is a confidential document based on attribute information associated with the second symbolically compressed document image, the security module being configured to prompt a user to enter authorization information before permitting output of the first symbolically compressed document image if the second document image is a confidential document and if the first and second document images match.

29. The document processing system of claim 28 further comprising a printer configured to receive a signal from the security module indicating whether to print a decompressed version of the first symbolically compressed document image.

30. The document processing system of claim 28 further comprising an transmission module to receive a signal from the security module indicating whether to transmit the first symbolically compressed document image.

31. The document processing system of claim 26 wherein the data processing system further comprises an monitoring module to determine if the second symbolically compressed document image is a copyrighted document.

32. The document processing system of claim 31 wherein the monitoring module is configured to automatically charge a copyright license fee for output of a decompressed version of the first symbolically document image if the second symbolically compressed document image is a copyrighted document and if the first and second symbolically compressed document images match.

33. An apparatus comprising a deciphering module to apply a hidden Markov model to decipher a sequence of template identifiers in a symbolically compressed document to recover a text string from the symbolically compressed document.

34. A method of extracting n-grams from a text, the method comprising:
automatically selecting alphabetic characters in the text that satisfy a predicate; and
concatenating the selected alphabetic characters to form n-grams.

35. The method of claim 35 wherein the predicate is a condition that the selected alphabetic characters follow respective spaces in the text.

36. An article of manufacture including one or more computer-readable media that embody a program of instructions to generate a text string from an input document image represented by a sequence of template identifiers for the purpose of reducing storage consumed by the input document image, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to replace the template identifiers with alphabet characters according to language statistics to generate a text string representative of text in the input document image.

37. The article of claim 36 wherein the one or more computer-readable media include one or more non-volatile storage devices

38. The article of claim 36 wherein the one or more computer-readable media include a propagated data signal.

39. The article of claim 36 wherein the program of instructions, when executed by the one or more processors in the processing system, causes the one or more processors to extract n-gram indexing terms from the text string by selecting alphabet characters in the text string that satisfy a predicate and then combining the alphabet characters in n-grams, n being an integer.

40. An article of manufacture including one or more computer-readable media that embody a program of instructions to generate a text string from an input document image represented by a sequence of template identifiers for the purpose of reducing storage consumed by the input document image, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to using a hidden Markov model to solve a substitution cipher formed by the sequence of template identifiers.